MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2018/2019

TDS2101 – INTRODUCTION TO DATA SCIENCE

(All sections / Groups)

26th OCTOBER 2018
9.00 AM – 11.00 AM
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 4 pages with 4 Questions only excluding the cover page

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) What are the **THREE** Big Data technologies used to support storage and analysis of Big Data? [3 marks]

(b) Describe **FOUR** A's roles of data scientist. [4 marks]

(c) Explain Big Data Veracity. Why is it a challenge for data scientists in both academia and industry to work with veracity of Big Data? [1 + 2 marks]

QUESTION 2

(a) The following data represents the readings of "bad" cholesterol levels for a group of senior citizens in an old folk's home.

0.72, 0.85, 1.09, 1.24, 1.37, 1.4, 1.47, 1.51, 1.58, 1.63, 1.64, 1.64, 1.67, 1.68, 1.69, 1.69, 1.7, 1.74, 1.75, 1.75

(i) Construct a box plot that represents the data above. [4 marks]

(ii) Identify mild and extreme outliers in the data. [1 mark]

(b) Describe **TWO** advantages of sampling a set of data from the actual data set. [2 marks]

(c) Describe stratified sampling method with an example. [3 marks]

CONTINUED...

QUESTION 3

(a) The following is a set of association rules and measures generated from applying association rule algorithm to a grocery database. For each of the following association rule with measures, interpret the rule and based on the interpreted result, identify **ONE** possible action that can be taken by the grocery shop owner.

(i) Product-A → Product-B (support = 0.9, confidence = 0.9 and lift = 2.0)
[2 marks]

(ii) Product-A → Product-B (support = 0.1, confidence = 0.2 and lift = 0.2)
[2 marks]

(iii) Product-A → Product-B (support = 0.1, confidence = 0.1 and lift = 2.0)
[2 marks]

(b) Using examples, describe **ONE** advantage and **ONE** disadvantage of column-oriented database. [4 marks]

QUESTION 4

Figure 1 shows the scores obtained by a group of Diploma students for three subjects namely Statistics, Programming and Database. These records are stored in a comma separated value file named *scores.csv*.

Figure 1: scores.csv

```
StudentID, Name, Majoring, Statistics, Programming, Database
1, Candy, Engineering, 90, 62, 55
2, Ahmad, Multimedia, 55,59,49
3, Ali, IT, 40, 75, 80
4, Lydia, Management, 75, 80, 68
5, Eleen, Engineering, 63,73, 78
6, Bernard, Communication, 82, 44, 20
7, John, IT, 60, 59, 55
8, Mary, Communication, 54, 62, 60
9, Edmund, IT, 65, 90, 40
10, Janice, Multimedia, 49, 51, 50
```

CONTINUED...

(a) Using R, create a function with no argument list and assign it to variable *firstfunction.* Assume that R's working directory is set to the right location to retrieve the file and relevant packages have been loaded into current R working environment. This function loads the scores.csv to R working environment and performs the following two operations. First, it presents the total number of students (in numbers) by Majoring. Then it displays the list of students who have scored above 50 marks for all the three subjects. The desired output is given as in Figure 2. [3 marks]

Figure 2: The desired output for Question 4(a)

```
>
> firstfunction()

 Communication     Engineering              IT      Management      Multimedia
             2               2               3               1               2
  StudentID   Name       Majoring Statistics Programming Database
1         1  Candy    Engineering         90          62       55
2         4  Lydia     Management         75          80       68
3         5  Eleen    Engineering         63          73       78
4         7   John             IT         60          59       55
5         8   Mary  Communication         54          62       60
>
```

b) Using R, create a function with no argument list and assign it to variable *secondfunction.* Assume that R's working directory is set to the right location to retrieve file and relevant packages have been loaded into current R working environment. This function loads the scores.csv to R working environment and perform the following operations. First, it replaces students' scores which are less than 50 with the word "FAIL". Secondly, once the replacement is completed, the students' records are displayed. Lastly it displays the total number of students that fail at least one subject. The desired output is as given as in Figure 3. [5 marks]

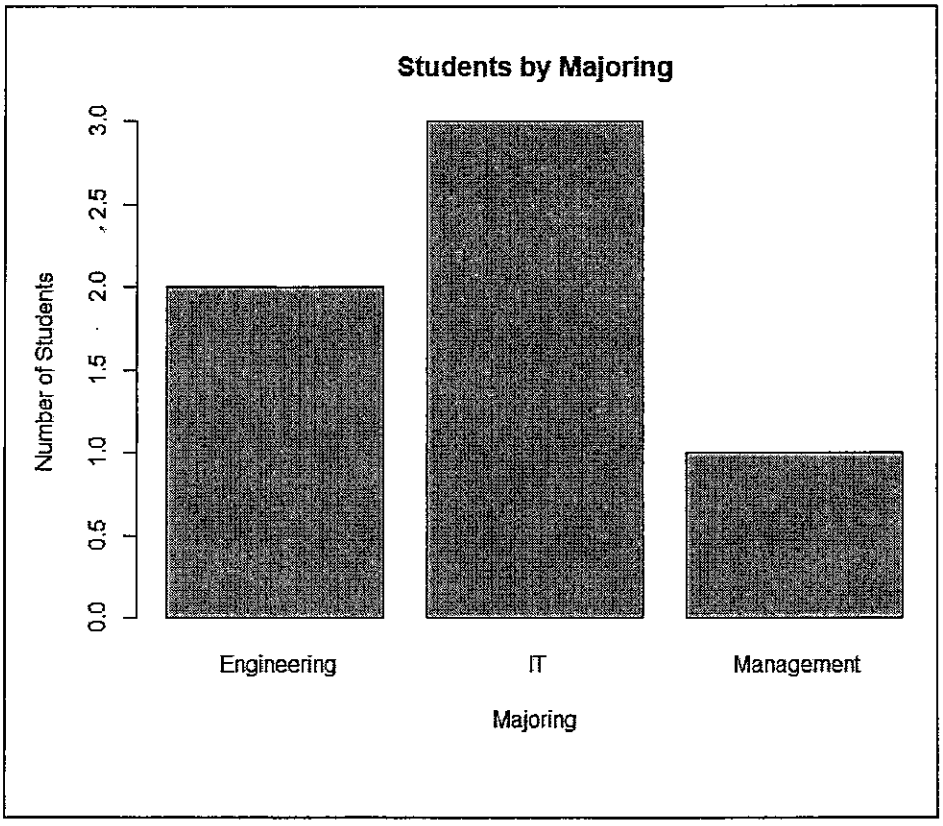
Figure 3: The desired output for Question 4(b)

```
>
> secondfunction()
   StudentID     Name       Majoring Statistics Programming Database
1          1    Candy    Engineering         90          62       55
2          2    Ahmad     Multimedia         55          59     FAIL
3          3      Ali             IT       FAIL          75       80
4          4    Lydia     Management         75          80       68
5          5    Eleen    Engineering         63          73       78
6          6  Bernard  Communication         82        FAIL     FAIL
7          7     John             IT         60          59       55
8          8     Mary  Communication         54          62       60
9          9   Edmund             IT         65          90     FAIL
10        10   Janice     Multimedia       FAIL          51       50
[1] 5
>
```

CONTINUED...

(c) Using R, write the code to build a bar plot that shows the total number of students by selected majoring i.e., Engineering, IT and Management. The desired output is given as in Figure 4. [2 marks]

Figure 4: The desired output for Question 4(c)



END OF PAGE.